JUDGEBERT: Assessing Legal Meaning Preservation Between Sentences

David Beauchemin[†], **Michelle Albert-Rochette**[‡], **Richard Khoury**[†] and **Pierre-Luc Déziel**[‡]
Université Laval, Québec, Canada
Computer Science Department[†] and Faculty of Law[‡]
david.beauchemin@ift.ulaval.ca, michelle.albert-rochette.1@ulaval.ca
richard.khoury@ift.ulaval.ca, pierre-luc.deziel@fd.ulaval.ca

Abstract

Simplifying text while preserving its meaning is a complex yet essential task, especially in sensitive domain applications like legal texts. When applied to a specialized field, like the legal domain, preservation differs significantly from its role in regular texts. This paper introduces FrJUDGE, a new dataset to assess legal meaning preservation between two legal texts. It also introduces JUDGEBERT, a novel evaluation metric designed to assess legal meaning preservation in French legal text simplification. JUDGEBERT demonstrates a superior correlation with human judgment compared to existing metrics. It also passes two crucial sanity checks, while other metrics did not: For two identical sentences, it always returns a score of 100%; on the other hand, it returns 0% for two unrelated sentences. Our findings highlight its potential to transform legal NLP applications, ensuring accuracy and accessibility for text simplification for legal practitioners and lay users.

1 Introduction

Automatic text simplification (ATS) aims to creates easier-to-read text while keeping the original meaning (Saggion, 2017). Evaluating whether a simplified text preserves the meaning of the original complex one is not trivial. Yet, it is critical for ATS and many other natural language processing (NLP) tasks, such as machine translation (Gatt and Krahmer, 2018). Evaluation of ATS is based on three dimensions of system generations: "fluency", "simplicity" and "meaning preservation". Fluency measures grammatical correctness, simplicity estimates how easy-to-understand the text is, while meaning preservation measures how well the output text's meaning corresponds to the original (Saggion, 2017). It is typical to use automatic metrics to assess these evaluations, such as BLEU (Papineni et al., 2002) and SARI (Xu et al., 2015), which tend to focus on only one of the three dimensions. For example, BLEU is commonly used to evaluate fluency, while SARI is for simplicity. More recent automatic NLP metrics use Transformer architecture to compute the ATS. For example, MeaningBERT (Beauchemin et al., 2023) is an evaluation metric that uses a fine-tuned BERT Transformer model to assess meaning preservation.

When applied to specialized fields, such as the legal domain, preservation differs significantly from its role in regular texts and has the potential to significantly impact all stakeholders. Inaccurate ATS can mislead users, cause legal issues, or represent a risk for the company that deploys the system (Štajner, 2021). A user can interpret an automatically-simplified text in a way that would not hold in court, creating a "legal gap" between the texts' meanings. For example, in 2024, an Air Canada passenger was misled about the airline's rules for bereavement fares when the company's AI chatbot hallucinated an answer inconsistent with their policies. The Tribunal found Air Canada guilty of "negligent misrepresentation" in this situation (Moffatt *v. Air Canada*, 2024). None of the metrics currently available specifically assess legal meaning preservation and have not been benchmarked against human judgment for this sensitive task. Developing a metric to assess whether a simplification still conveys the same legal meaning is crucial to minimizing risk in legal NLP applications. To this end, we introduced "legal meaning" as a substitute to "meaning preservation" to evaluate ATS system output for legal text simplification (TS). Our three contributions are:

1. We proposed a new dimension to evaluate ATS system output for legal ATS;
2. We proposed FrJUDGE[1], a **Fr**ench corpus of insurance legal meaning **JUDGE**ments to assess legal meaning preservation between an insurance contract and its simplified text; and
3. JUDGEBERT, a new fine-tuned BERT metric de-

[1] https://github.com/GRAAL-Research/JUDGEBERT

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 92–118
November 4-9, 2025 ©2025 Association for Computational Linguistics

signed to assess the legal meaning preservation between two French legal sentences, which we have trained to correlate with human judgment on French insurance text.

This paper is outlined as follows: first, we study the relevant ATS metrics research and corpora in Section 2. Then, we propose a definition of legal meaning, and we present our corpus in Section 3, along with our new trainable metric, JUDGEBERT in Section 4. To demonstrate its quality, we also present a set of experiments in Section 5, and, following Beauchemin et al. (2023), we will also conduct a set of sanity checks. Finally, we will discuss our results in Section 6 and conclude in Section 7.

2 Related Work

2.1 Human Evaluation and Automatic Metrics for Meaning Preservation

Since automatic metrics are a proxy for human judgments for ATS, they should correlate well with human ratings. However, Sulem et al. (2018) found low to no correlation between BLEU and meaning preservation dimensions when sentence splitting is involved, a typical simplification operation used notably for legal texts (Garimella et al., 2022). They also pointed out that BLEU is sensitive to the length of the compared texts and does not consider semantic variability between sentences that differ on synonymous words or in word order.

Since word-embeddings-based metrics can better account for semantic variability between sentences (Zhang et al., 2019), Beauchemin et al. (2023) have conducted a correlation analysis on meaning preservation of 22 ATS metrics. These include popular non-Transformer and Transformer ones. Their results show that many non-Transformer metrics correlate poorly with human judgment, and most Transformer ones correlate weakly. In addition, they also conducted benchmarking tests to evaluate meaning preservation between pairs of identical and unrelated sentences. These tests show that many automatics metrics fail even in these simple tasks. Furthermore, they proposed MeaningBERT, a fine-tuned Transformer-based metric that correlated better with human judgment and passed the benchmarking tests. Nevertheless, none of these metrics focus on legal meaning.

2.2 French Legal Text Simplification Datasets

Only three TS French datasets are available in the literature, and none focus on legal documents (Ryan et al., 2023). Indeed, Alector (Gala et al., 2020) focuses on literacy and scientific texts, while CLEAR (Grabar and Cardon, 2018) on medical text, and WikiLarge FR (Cardon and Grabar, 2020) on informative text from Wikipedia. None of these available corpora are suited to our needs since legal documents differ from other texts: they are lengthier and use specialized vocabulary (Katz et al., 2023). Only two corpora of legal documents are available in French: RISCBAC (Beauchemin and Khoury, 2023), a set of synthetic bilingual automobile insurance legal contracts, and EUR-Lex-Sum (Aumiller et al., 2022), a multi- and cross-lingual set of summaries of legal acts from the European Union law platform. However, neither dataset includes simplifications or human annotations.

3 FrJUDGE: a French Corpus of Insurance Legal Meaning Judgments

In this section, we introduce the **Fr**ench corpus of insurance legal meaning **JUDG**m**E**nts (FrJUDGE), which is the first legal meaning judgment dataset in any language. FrJUDGE consists of 297 human-annotated French sentences taken from property damage insurance forms used by two insurance regulators, namely the *Bureau d'assurance du Canada* (BIC, 2009), and the *Autorité des marchés financiers du Québec* (AMF, 2014). As illustrated in Table 1, each dataset instance consists of a legal sentence, a simplification, and human annotations (simplicity, characterization and legal meaning). Both sources are publicly available online, and we obtained authorization to publish them under a CC-BY 4.0 license.

3.1 Legal Meaning

We argue that "meaning" and "legal meaning" differ because, for typical ATS, synonyms can be used to convey the same meaning, while for legalese, synonyms do not necessarily convey the same meaning. For example, in the common language, "automobile" and "vehicle" share the same meaning. However, for legalese, the first means any vehicle moved by a "mechanical force" and the latter by "mechanical or human force" (Quebec, 2022a). Thus, an automobile is a vehicle, but a vehicle is not necessarily an automobile. For example, a bicycle is a vehicle that fits the description of a vehicle but not an automobile.

Given that no previous work nor automatic metric focuses on legal meaning and that our goal is

Legal Sentence	Simplified Sentence	Simplicity Level	Characterization	Legal Meaning
L'assuré désigné est le propriétaire réel et le titulaire de l'immatriculation du véhicule désigné.	*L'assuré désigné possède le véhicule désigné. Il détient aussi son immatriculation.*	*Aussi simple à lire*	2	8

Table 1: Example of an instance from FrJUDGE containing a legal sentence and human annotations (simplification, simplicity level, characterization and legal meaning).

to determine whether or not ATS systems can simplify texts while maintaining their meaning from a legal standpoint, we must rigorously define what "legal meaning" is. Only a few articles focus on legal TS, and none specifically study the preservation of legal meaning between two texts. However, Hagan (2023) proposes 22 actionable criteria for legal question-answering that any legal AI system should be benchmarked on to fully assess its capabilities and limitations, and guide policymakers and regulators. These criteria are closely related to "how a professional lawyer should conduct themselves in their practice". Two of these criteria are particularly interesting for our work: a "response is robust and comprehensive, covering details and exceptions" and a "response does not misrepresent the substantive law". Using these two criteria, we proposed the following definition for "legal meaning" as a metric to assess the quality of a legal ATS system: "**Legal meaning measures how well the output text conveys the legal details and exceptions and does not misrepresent the law**".

3.2 FrJUDGE Corpus

3.2.1 Data Collection

Sentences in FrJUDGE were collected manually from the two insurance forms. Specifically, we examined all sentences and extracted 312 text blocs based on three criteria: text blocs are

1. between 1 and 5 sentences long;
2. not boilerplate texts such as a title; and
3. college-level reading level grade ($\leq$ 50) on the French Flesch-Kincaid grade level (FKGL) (Kandel and Moles, 1958) to focus on more challenging sentences in an insurance contrat.

For example, the sentence (translated) "The city and province of the address written in this section 1 constitute the designated vehicle's principal place of use, storage and parking." passes the first two criteria. However, it scored 69.87 on the FKGL, so it was not selected.

3.2.2 Automatic Text Simplification

Since few ATS systems exist in French and none are designed explicitly for legal texts, no pretrained models are available to generate French ATS. Thus, all sentences in the corpus were automatically simplified using the OpenAI GPTs model through their API. We selected this approach since it has been shown by Feng et al. (2023); Kew et al. (2023); Wu and Arase (2024) that foundational large language models (LLMs) generate less erroneous simplification outputs than state-of-the-art approaches; thus, they are effective ATS systems, even when using zero-shot prompting. Moreover, it also has been shown by Nozza et al. (2023); Madina et al. (2024) that GPTs are effective ATS systems in languages other than English, such as Italian and Spanish. We present the details used for generation in Appendix A and examples in Appendix B.

3.2.3 Human Evaluation Methodology

Following the arguments of van der Lee et al. (2019), we present our human evaluation methodology's in this section.

Selected models. We selected GPT4-turbo with zero-shot prompting.

Number of outputs. We randomly selected 297 instances for annotation and 15 for practice[2].

Presentation and interface. We used a customized version of the Prodigy annotation tool (Montani and Honnibal, 2018), and we present in Appendix C the interface (in French). Annotators use our annotation procedure to annotate each instance randomly. Like the ATS system, annotators were not given the overall legal documents.

Annotators. We selected five native French-speaking law students at the Faculty of Law of University Laval as our annotators. A meeting was held with them to introduce the task, instructions, and annotation guide and interface. Instructions included that they must spend at most 5 minutes per sentence pair. Furthermore, 15 instances were annotated during a pilot phase to familiarize them with the task. Finally, during a second meeting after evaluating the practice instance, annotators received feedback and advice on what phenomena

[2] These practice annotations have been used to help annotator practice their tasks and adjust our guidelines; these examples have been discarded from the final dataset.

they should be cautious about. Recognizing the significant contribution of our annotators, they were remunerated fairly according to the University's hourly salary pay scale. Each annotator completed their work in at most 30 hours. We provide in-depth details of the evaluation setup in our Human Evaluation Datasheet (Shimorina and Belz, 2021) in Appendix F.

Legal Meaning Scale. Given that we have previously defined what the term "legal meaning" means, we now define "**legal meaning metric**" as the metric that measures the legal meaning between the legal original and the simplified text. We will refer to this metric as the `legal meaning preservation` (`LMP`). To this end, we designed a Likert scale (Likert, 1932) ranging from one to ten. A simplified text that scores a ten means that the legal meaning between the two texts "tends to be preserved"[3]. On the other hand, one receiving a score of 1 does not match the original legal meaning at all.

Annotation Procedure. The annotators must determine one of TS's three dimensions, simplicity, along with our new fourth dimension to replace the "meaning" dimension. We choose not to evaluate fluency since Wu and Arase (2024) have shown that GPT4 fluency capabilities are near perfect (2.98/3).

First, the annotators assess the `simplicity` of the simplified text. We have adopted a simpler version of the eight-level ordinal scale proposed by Primpied et al. (2022), which uses intuitive perception levels of text difficulty ranging from children's stories (lowest) to legal documents (highest). Our initial pilot found the scale to be too complex for our case, which is coherent with Stodden (2021) conclusion that "interpretation of the simplicity scale is consistent when rated by experts [...]". Indeed, we found that our annotators tended to assign mostly either the "legal documents" level or a level below this. Moreover, annotators expressed their concern about the scale, which motivated us to change it.

Thus, our version uses four levels (translated): "Easier to read", "Equal to read", "More difficult", and "No simplification"[4]. Since our annotators are legal experts, we selected this approach because expert annotators tend to "inject their own opinions and biases" during annotations (van der Lee et al., 2019).

Second, following the work of Garneau et al. (2022), the annotators use a three-step process to assess an instance's `LMP`. First, they decide the `characterization` of the text. Characterization refers to qualifying laws (Fréchette, 2010)[5]. For example, risks that are not covered in an insurance contract, such as nuclear damage, are characterized as "exclusions or restrictions". Each annotated sentence is characterized into one of our 18 classes detailed in Appendix D. This step helps annotators identify the type of legal text the instance refers to; it does not impact the `LMP` score. With this approach, our annotators can rely on their legal background and education to assess whether a simplification respects a class's characterization elements, such as whether it states a proper definition that respects Quebec legislation. In the second step, the annotators assign a preliminary `LMP` score. Garneau et al. (2022) observed that legal experts naturally divide their decisions into three regions instead of directly assessing a score between 1 and 10. Consequently, following their work, we split our legal accuracy scale into the three score brackets listed below.

7 - 10 – Accurate. Means the simplification seems to entail the legal details and exceptions properly and does not misrepresent the law; it is considered to "tends to be" accurate.

2 - 6 – Seems Imprecise. Means the generation seems to improperly entail the legal details, exceptions and slightly misrepresents the law.

1 – Off-Track. Means the simplification is obviously erroneous, does not entail the legal details and exceptions, and misrepresents the law.

Once the annotators have chosen the score bracket where the simplification belongs, they move on to the final step: looking for legal errors in the output. We identify four types:

- **Hallucinations** are facts the model generates despite not appearing in the original text. For example, the simplified text might specify that the insured is covered for a particular risk, while the original clause does not mention it.

[3] Since most of our annotators were reluctant to state that the two sentences were equivalent, and due to the legal risk of stating that a sentence is "perfectly preserving its legal meaning", we choose to be less assertive in our scale.

[4] "No simplification" applied to the case where the simplification is identical to the original text or in another language.

[5] It is worth mentioning that since the sentence is isolated, it can be challenging to select the proper characterization and sometimes more than one can apply. For the latter, annotators must select the one that seems to apply the most.

- **Omissions** occur when essential facts are in the original text but are not in the simplified text generated by the model. For example, the original clause might specify a maximum coverage amount, but the simplified text does not.
- **Consistency** issues occur when the model simplifies a juridical term but does not use the simplified term for all occurrences of the juridical term. For example, if the original clause refers to an "automobile" and the simplification replaces it with "vehicle", we do not consider it a consistency error. On the other hand, if the simplification alternates between using "automobile" and "vehicle", it would be an error.
- **Confusions**: factual mistakes characterized by mismatches between the source and the generation. For example, the source says that the insured must declare claims as soon as possible, but the generation states otherwise.

Each error reduces the output's score by one point, starting from the bracket's maximum. The output's score can never drop below the score one (1). To summarize this process, Figure 1 conceptualizes the corresponding Likert scale.

3.2.4 Annotation Results

We provide in Figure 2 the breakdown, by annotator, of all annotation criteria. First, we can see in Figure 2a and Figure 2b that `simplicity level` and `characterization` are distributed similarly among all annotators, except for the annotator E. Indeed, annotator E finds more frequently simplified text easier to read than the other annotators and assigns most of its `characterization` annotations into the first class (i.e. description of endorsement). Since this class can act as a generic class, this characterization is adequate but could be more precise. However, since this step acts as an intermediary one, it does not negatively affect the quality of the annotations. Second, we can also see that for `LMP`, we seem to have two clusters of similar annotation distributions. Annotators A, B and C generally assign similar scores to each other, while annotators D and E often behave similarly to each other but differently from the first group. Furthermore, the first group has a higher frequency of perfect scores (10). On the other hand, annotators D and E more frequently attribute "Off-Track" score (1), meaning they were more strict in their initial reading of the simplification. This could be due to the annotators' domain expertise, which allows them to infer the possible context and case law, which was not available for the ATS system. Nevertheless, since `LMP` is subject to the legal counsellor's interpretation, this situation is not problematic as it reflects the complexity of the task.

	Agreement (%) (↑)	Krippendorff's α (↑)	Accuracy (%) (↑)
`Simplicity Level`	57.17	0.18	48.11
`Characterization`	60.24	0.55	58.05
`Legal Meaning Preservation`	25.96	0.10	18.48
Average	47.74	0.28	42.84

Table 2: Annotators inter-agreements metrics per annotation task and average. ↑ means higher is better.

Since we have multiple annotators, we present in Table 2 the inter-agreement statistic of our annotators[6], namely the percent agreement, the Krippendorff's alpha coefficient (KAC) (Hayes and Krippendorff, 2007) and the accuracy score to measure inter-annotator agreement. We can see that annotators have a high agreement over the agreement score and accuracy for the `simplicity level` and `characterization`. However, the KAC of the `simplicity level` and `LMP` shows a weak agreement. This is because annotators D and E regularly disagree with the other three annotators.

Final Annotation To select the final annotation, we use a majority vote for `simplicity level` and `characterization`, and in case of ties we randomly select between the equal options. For the `LMP`, we compute an average score.

3.3 Corpora Analysis

Table 3 presents some key statistics of FrJUDGE and the other French simplification and legal corpora introduced in Section 2.2, where the lexical richness corresponds to the ratio of a sentence number of unique words over the overall vocabulary cardinality without removing the stop words or normalizing them (Van Hout and Vermeer, 2007). We excluded Alector since that dataset is not available for download. For all corpora, we have used the latest official version on the HuggingFace Datasets Hub. All statistics were computed using SpaCy (Honnibal et al., 2020) and exclude new lines (`\n`), whitespaces and punctuations. We can see in Table 3 that our FrJUDGE datasets' statistics, namely lexical richness and length size, are quite similar to those of the compare corpora.

[6] Computed using the Prodigy inter-annotator Agreement Python package toolkit (Montani and Honnibal, 2018).

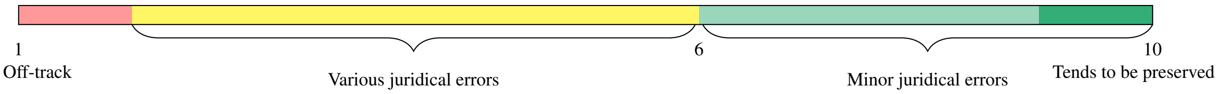
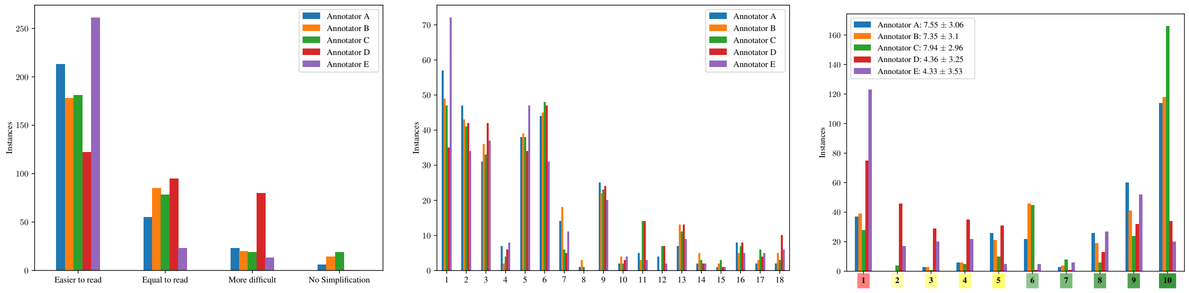


Figure 1: The four-section Likert legal meaning scale used for annotation. The annotators first decide where the generation sits between the four regions. Then they remove points for every error encountered.



(a) Distribution, per annotator, of the `simplicity level` annotation.

(b) Distribution, per annotator, of the `characterization` annotation.

(c) Distribution, per annotator, of the `legal meaning preservation` annotation and the average score and one standard deviation.

Figure 2: Distribution, per annotator, of the annotation for all three aspects.

	FrJUDGE		CLEAR		WikiLarge FR	
	Complex	Simple	Complex	Simple	Complex	Simple
# of sentences	297	297	4,596	4,596	297,753	297,753
Lexical richness	1.2	1.1	2.4	2.28	0.59	0.5
Avg sent len	18.83	14.24	19.65	19.50	14.57	11.73

Table 3: Aggregate statistics of FrJUDGE, and the French simplification introduced in Section 2.2.

4 JUDGEBERT

We propose JUDGEBERT, the first supervised automatic metric for LMP that correlates with human judgment and passes the two sanity checks. JUDGEBERT is built upon the CamemBERT-baseV2 model (Antoun et al., 2024), but uses a regression head instead of a classification one and feeds sentences pair into the network by concatenating them with a [SEP] token. CamemBERTV2-base is the smallest CamemBERTV2 model, built on the RoBERTa architecture with 112 million parameters. It comprises 12 768-transformer layers and attention heads. JUDGEBERT is trained by fine-tuning the pretrained model for at most 100 epochs with an initial learning rate of $5e-5$, a patience of 5 epochs, a batch size of 16, and a linear learning rate decay as suggested by Mosbach et al. (2021) using a 10-fold approach with a different random seed ($[42, \cdots, 51]$) to split the dataset in a 60-10-30 % train-validation-test split and initialize the new regression attention head weights.

5 Experimental Setup

In this section, we discuss our experimental setup. First, we discuss the automatic metrics we studied and the two sanity checks used in our experiments and finish with the training details of JUDGEBERT.

5.1 Selected Metrics

Since no automatic metric exists for preserving meaning in French or legal meaning, our experimentation builds upon previous studies on meaning preservation. Specifically, we rely on the findings of Beauchemin et al. (2023), which suggest that Transformer-based metrics correlate better with human judgments. Thus, for our experiments, we limited ourselves to Transformer-based metrics. We selected the following ones:

- **BERTScore** (Zhang et al., 2019) uses BERT monolingual English contextual word embeddings and computes the cosine similarity between the tokens of two sentences. It can compute precision, recall, and F1 Score over two sentences. We selected only the F1 Score since Beauchemin et al. (2023) have shown that BERTScore precision and recall scores are relatively similar for meaning preservation, and our initial experiments have shown similar results.
- **Sentence Transformer** (Reimers and Gurevych, 2019) uses a siamese network to compare two-sentence embeddings using the cosine similarity. We selected the best monolingual English pretrained (`sentence-t5-xxl`) (SBERT) and multilingual (`distiluse-base-multilingual-cased-v1`) (SBERT-Multi) model from the official Python library.

- **Coverage** was introduced by Laban et al. (2020) to assess the meaning preservation between two texts. It uses a cloze test (Taylor, 1953) to assess whether a monolingual English LM can fill the masked source document using a summary generated from it.
- **QuestEval** (Rebuffel et al., 2021) is a metric designed to evaluate ATS output quality using synthetic questions and a monolingual English QnA model to respond to the generated questions using the simplification. The intuition is that if a simplified text conveys the same information as the source, a QnA model should be able to respond appropriately to a set of questions based on the source text.
- **LENS** (Maddela et al., 2022) is a trained metric for ATS quality assessment built upon monolingual English BERT.
- **MeaningBERT** (Beauchemin et al., 2023) is a trained metric built upon a monolingual English BERT-like model for meaning preservation between two sentences, but it does not focus on legal meaning.

5.1.1 Nomalization

We normalize the outputs of the different systems by decimal scaling, so those whose outputs are in $[0, 100]$ or $[0, 1]$ all line up in a $[0, 10]$ range.

5.1.2 Semantic Capabilities

It is essential to note that only one metric utilizes multilingual embeddings, while all the others rely on monolingual English embeddings. Thus, none are specialized in French, while ours leverage French-specialized embeddings. Nonetheless, our objective is to study relevant metrics and evaluate if they correlate well with human judgment, regardless of their semantic initial capabilities.

5.2 Sanity Checks

As per Beauchemin et al. (2023), we also conduct two automated sanity checks as an alternative evaluation of the metrics. The checks evaluate LMP between identical and unrelated sentence pairs. In these checks, the legal meaning preservation is a non-subjective measure that does not require human annotation for its assessment. They are trivial and minimal thresholds that a good automatic LMP metric should be able to achieve. For our experiments, we compute the ratio of identical sentence pairs that score equal or greater to 99%, and the ratio of unrelated sentence pairs that score equal or below 1%. We allow a 1% margin in each case to account for computer floating-point inaccuracy. To generate unrelated sentences, the authors of Beauchemin et al. (2023) used GPT-2 to generate a random sentence and pair it with an unannotated sentence taken from the ASSET (Alva-Manchego et al., 2020) corpus. This approach works well for sentences that use common language. However, our legal corpus uses less common vocabulary than standard NLP corpus. Pairing sentences from legal documents with unrelated randomly generated sentences could make the sanity checks too trivial. Instead, we sampled a sentence from the Québec Automobile Insurance Act (Quebec, 2022b) and matched it with a sentence taken from the Québec Road Safety Code (Quebec, 2022a) that reached a maximum ROUGE-$[1, 2, \text{L}]$ and BLEU score of 0.25 and 25, respectively. Table 4 illustrates an example of two matched sentences to illustrate how the two sentences use similar lexical vocabulary yet are unrelated.

5.3 Training and Evaluation Datasets

Since JUDGEBERT is a trainable metric, we specify the datasets used to benchmark all metrics.

5.3.1 JUDGEBERT Training Datasets

To train JUDGEBERT, we use FrJUDGE legal meaning human annotations, and the complex and LLM-generated simplification sentences to form a triple. During training, we use two datasets: one using FrJUDGE 297 sentence triplets and a second that uses 594 sanity-check data augmented (DA) sentence triplets along with the FrJUDGE corpus, for a total of 891 sentence triplets. We will refer to them as JUDGEBERT and JUDGEBERT-DA, respectively[7]. We hypothesize that our data augmentation approach will improve JUDGEBERT's performance on our two sanity checks, thus creating a more logical response by the metric for such cases.

5.3.2 Evaluation Datasets

We evaluate all selected metrics and JUDGEBERT on the same FrJUDGE test split during the test phase, either using the split with or without data augmentation (DA).

5.3.3 Evaluation Metrics

To investigate how well metrics correspond with human judgments of LMP, we evaluate them as ma-

[7] All corpora use human-annotated sentence triplets; the data augmented corpus only adds a new sentence to the corpus.

The persons referred to in sections 97, 99 and 100 must, at the request of a peace officer, surrender their permit for examination.
The Minister of Revenue may, without the consent of the person concerned, communicate to the Company any information necessary for the administration of the International Registration System.

Table 4: Example of two matched unrelated sentences (translated) randomly sampled from two legal sources. The pair reach at most a ROUGE-[1, 2, L] and BLEU scores of 0.25 and 25, respectively.

chine learning models. We use the Pearson correlation (Zar, 2005) and RMSE (James et al., 2013) between each metric's scores and human judgment.

5.3.4 Sanity Checks Hold-out Datasets

To benchmark all metrics and JUDGEBERT on our sanity checks, we use a hold-out dataset composed of unseen sentences taken from the unused sentences in our two legal-related corpora to create an unrelated match and generate 297 related and unrelated sentences as a hold-out evaluation corpus.

6 Metrics Ratings Analysis

In this section, we analyze the selected metrics and JUDGEBERT for their ability to evaluate LMP. Table 5 presents the evaluation results of all metrics and our two sanity checks. For JUDGEBERT, we display the average score and one standard deviation. **Bolded** values are the best results per column. We also display in Appendix E the training and evaluation loss for JUDGEBERT and discuss overfitting risk.

DA	Metric	Pearson (↑)	RMSE (↓)	% > 99% (↑)	% > 1% (↑)
False	BERTScore	0.46	3.61	**100.00**	0.00
	Coverage	0.19	2.82	0.00	0.00
	LENS	0.38	2.57	0.00	0.67
	MeaningBERT	0.17	3.51	**100.00**	0.67
	QuestEval	-0.05	2.99	0.00	0.00
	SBERT	0.13	3.25	**100.00**	0.00
	SBERT-Multi	0.06	3.35	0.00	0.00
	JUDGEBERT	0.74 ± 0.02	1.72 ± 0.10	0.00	0.00
True	BERTScore	0.94	5.09	**100.00**	0.00
	Coverage	0.90	2.20	0.00	0.00
	LENS	0.56	3.87	0.00	0.67
	MeaningBERT	0.81	3.98	**100.00**	0.67
	QuestEval	0.68	3.82	0.00	0.00
	SBERT	0.92	2.84	**100.00**	0.00
	SBERT-Multi	0.90	2.39	**100.00**	0.00
	JUDGEBERT-DA	**0.97 ± 0.00**	**1.01 ± 0.07**	**100.00**	**100.00**

Table 5: Results of the selected metrics and JUDGEBERT trained with or without data augmentation (DA). We also present one standard deviation for trained models. **Bolded** values are the best results overall. ↑ means higher is better, while ↓ mean otherwise.

6.1 Metrics Ratings and Human Judgments

First, we can see in Table 5 that Pearson correlation scores vary greatly between metrics, with an average correlation between $[-0.05, 0.74]$ and $[0.56, 0.97]$, with and without DA, respectively. This shows that not all metrics are suitable for our task. Indeed, we can see that most selected metrics have a low to moderate degree of correlation with human judgment, with BERTScore reaching the second-highest score. We can also see that all metrics achieve a higher correlation with human judgment when DA is introduced, meaning they can, to a certain degree, be compliant with our two sanity checks. Furthermore, JUDGEBERT achieves the highest correlation with human judgment, with a near-perfect correlation when trained with DA.

On the other hand, we can see that all selected metrics achieve poor performance on RMSE, higher than JUDGEBERT with and without DA. Since our labels are on a 10-point Likert scale, the RMSE corresponds to the number of levels of difference between the model's output and human judgement. Our results thus demonstrate that the selected metrics are, on average, very different from human judgments. Furthermore, since we want to assess LMP, the impact of a "close enough" score differs depending on whether the score is higher or lower than the human evaluation. Indeed, in practice, a system that undershoots human judgment is simply strict in the simplifications it accepts, but one that overshoots human judgment is unacceptably permissive of bad simplifications. Thus, we present in Table 6 the percentage of predictions with a higher output than the human judgment on the corpus without DA. For all metrics, except our JUDGBERT models, the output score is regularly higher than human judgments. It shows that other metrics are inadequate for LMP.

6.2 Metrics Sanity Checks

We can see in Table 5 that only three metrics always return the expected value of 100% (e.g., 99% to account for rounding error) when comparing two identical sentences: BERTScore, SBERT, SBERT-Multi and MeaningBERT. These results are expected for all metrics, as BERTScore employs an algorithm that returns a perfect score when the two

Metric	% > labels (↓)
BERTScore	82.22
Coverage	27.78
LENS	30.00
MeaningBERT	82.22
QuestEval	27.78
SBERT	76.67
SBERT-Multi	77.78
JUDGEBERT	**0.00 ± 0.00**
JUDGEBERT-DA	**0.00 ± 0.00**

Table 6: Percentage of predictions with a higher rating than the human judgments of the selected metrics and JUDGEBERT on the test set without DA. **Bolded** values are the best results. ↑ means higher is better.

texts are identical. MeaningBERT was trained to do so, and SBERT and SBERT-Multi both use cosine similarity between embeddings to compute the similarity. Thus, two similar sentences will return the same vectors.

On the other hand, none of the metrics achieve a perfect performance on the second check. This poor performance is similar to the results observed by Beauchemin et al. (2023). These authors hypothesize that BERT-like metrics that use contextualized embeddings can hallucinate connections and common meaning between the two sentence vectors even when none exist, thus returning a non-zero rating. This is likely our case since we use unrelated sentences with a similar legal lexicon but from two different sources. It shows that without proper legal knowledge, unrelated sentences can seem similar. This is a significant limitation of existing metrics in our case: since we evaluate LMP, generating a score different from zero for two completely unrelated sentences significantly reduces a metric's credibility for a legal counsellor.

Finally, we can see that with DA, JUDGEBERT-DA can pass both sanity checks. It shows that an LM cannot capture the coherent logic embedded in our sanity checks without being given proper examples.

7 Conclusion and Future Work

This paper proposes a new metric to assess legal meaning preservation between two legal sentences, specifically in the context of text simplification. However, our metric could also be used for other tasks. We also proposed FrJUDGE, a new legal meaning judgment dataset consisting of 297 human-annotated sentences taken from French insurance legal documents. To demonstrate its quality and versatility, we compared our work against a set of Transformer-based metrics in the literature applied to FrJUDGE. Further, we applied two automatic sanity checks to evaluate meaning preservation between identical and unrelated sentences. In future work, we aim to study how JUDGEBERT generalizes to other languages and tasks. We also aim to increase FrJUDGE's size by including other insurance products, such as group insurance. Finally, we also want to expand FrJUDGE's size by including pieces of text that are not jurisdiction-specific, such as French versions of arbitration and mediation clauses, which are subject to international conventions and not region-specific.

Limitations

All the sentences included in FrJUDGE have been extracted from para-governmental official sources. Therefore, they are guaranteed to be meaningful, making FrJUDGE a challenging dataset. However, text instances are relatively short and are analyzed by legal experts outside their context; thus, this differs from how contracts are typically analyzed (i.e. as a whole), and the application of a contract depends to a large extent on the facts (Cardon and Grabar, 2020). Such an approach, which contextualizes the overall document for text simplification, is more coherent with the recent work of Agrawal and Carpuat (2024). However, doing such an evaluation would be more costly and complex to orchestrate. Nevertheless, such approaches have been conducted with corpus such as CUAD (Hendrycks et al., 2021). To generate such a complex dataset, the cost of CUAD is estimated to be in the millions of dollars, whereas our annotation budget was USD 5,000.

JUDGEBERT has been trained on a relatively small dataset (i.e. FrJUDGE) for such a large model, and it has only seen a subset of all types of legal documents (namely insurance text). Moreover, our trained models were not tested with an out-of-domain (OOD) split to assert any overfitting risk. Thus, JUDGEBERT may have overfitted our training splits. However, we hope that the NLP community's interest in this work will lead to the development of robust metrics to assess the legal aspect of deep learning models.

As shown in Section 3.2.4, assessing the legal meaning precision of text is complex and is subject to interpretation. Interpreting whether or not a reformulation of a text conveys the same legal meaning will always be an approximation, and the only real complete test would be to discuss it in tribunals. However, such assessments are nearly impossible on a large scale. Thus, we argue that our approach can give insightful information to any legal practitioner on the overall `legal meaning precision` of a legal TS rather than a complete juridical analysis. However, our approach should not be considered legal advice, and `JUDGEBERT` should not be considered a comprehensive legal expert.

Finally, the metrics we selected in our study (Section 5.1) are mostly English-based approaches, yet we applied them to French text, which may give our French-based approach a potentially unfair advantage. Thus, our study cannot conclude whether some of these metrics are irrelevant to the preservation of legal meaning in English.

Ethical Considerations

FrJUDGE may serve as training data for French legal classifiers (Batra et al., 2021), as an expert source for text to structure legal expert systems (Janatian et al., 2023), or for training specialized LLM in French (Douka et al., 2021; Garneau et al., 2021), which may benefit the quality of generated texts in the legal field (Tan et al., 2023; Kapoor et al., 2024). Our corpus can be used to enhance online legal resources, providing laypeople with access to juridical services (Hagan, 2023; Kapoor et al., 2024). We acknowledge that such text generation progress could lead to the misuse of LLMs for malicious purposes, such as legal disinformation or harmful text generation (Weidinger et al., 2021; Bender et al., 2021; Hagan, 2023; Kapoor et al., 2024). However, our corpus can also be used for training adversarial defence systems against such misuses and to train artificial text detection models, (Lewis and White, 2023; Kumar et al., 2023).

`JUDGEBERT` may serve as a metric for evaluating LLMs in the legal and insurance domains. Legal documents are more challenging to read than typical documents; simplifying these documents can prove to be costly, so assessing the quality of legal documents is also costly (Hendrycks et al., 2021). We acknowledge that using trained metrics could lead to misuse and blind faith in users who trust such metrics. Nevertheless, our metric can be further improved to increase laypersons' access to proper legal expertise.

Acknowledgements

This research was made possible thanks to the support of a Canadian insurance company, NSERC research grant RDCPJ 537198-18 and FRQNT doctoral research grant. We thank the reviewers for their comments regarding our work as well as our colleagues from the *Groupe de recherche en apprentissage automatique de l'Université Laval* who peer-reviewed our manuscript. Lastly, we wish to thank the annotators who evaluated the generations.

References

Sweta Agrawal and Marine Carpuat. 2024. Do Text Simplification Systems Preserve Meaning? A Human Evaluation via Reading Comprehension. *Transactions of the Association for Computational Linguistics*, 12:432–448.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A Dataset for Tuning and Evaluation of Sentence Simplification Models With Multiple Rewriting Transformations. In *Annual Meeting of the Association for Computational Linguistics*, pages 4668—-4679.

Autorité des marchés financiers AMF. 2014. Formulaire de police d'assurance automobile du québec. Accessed: 2024-05-12.

Wissam Antoun, Francis Kulumba, Rian Touchent, Éric de la Clergerie, Benoît Sagot, and Djamé Seddah. 2024. CamemBERT 2.0: A Smarter French Language Model Aged to Perfection.

Dennis Aumiller, Ashish Chouhan, and Michael Gertz. 2022. EUR-Lex-Sum: A Multi-and Cross-lingual Dataset for Long-form Summarization in the Legal Domain. *arXiv:2210.13448*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Soumya Batra, Shashank Jain, Peyman Heidari, Ankit Arun, Catharine Youngs, Xintong Li, Pinar Donmez, Shawn Mei, Shiunzu Kuo, Vikas Bhardwaj, Anuj Kumar, and Michael White. 2021. Building adaptive acceptability classifiers for neural NLG. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 682–697, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

David Beauchemin and Richard Khoury. 2023. RISC: Generating Realistic Synthetic Bilingual Insurance Contract. *Proceedings of the Canadian Conference on Artificial Intelligence*. Https://caiac.pubpub.org/pub/k18zu6c9.

David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. MeaningBERT: Assessing Meaning Preservation Between Sentences. *Frontiers in Artificial Intelligence*, 6.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the ACM conference on fairness, accountability, and transparency*, pages 610–623.

Bureau d'assurance du Canada BIC. 2009. Formulaire d'assurance habitation du québec. Accessed: 2024-05-12.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Rémi Cardon and Natalia Grabar. 2020. French Biomedical Text Simplification: When Small and Precise Helps. In *Proceedings of the International Conference on Computational Linguistics*, pages 710–716.

Vincent Caron. 2024. *Assurance, biens et responsabilité*. Édition Yvon Blais, Montréal. Préface de l'honorable Pierre J. Dalphond.

Stella Douka, Hadi Abdine, Michalis Vazirgiannis, Rajaa El Hamdani, and David Restrepo Amariles. 2021. JuriBERT: A Masked-Language Model Adaptation for French Legal Text. In *Natural Legal Language Processing Workshop*, pages 95–101. Association for Computational Linguistics.

Yutao Feng, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. Sentence Simplification via Large Language Models. *arXiv:2302.11957*.

Pascal Fréchette. 2010. La qualification des contrats: aspects théoriques. *Les Cahiers de droit*, 51(1):117–158.

Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C Ziegler. 2020. Alector: A parallel corpus of simplified french texts with alignments of misreadings by poor and dyslexic readers. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1353–1361.

Aparna Garimella, Abhilasha Sancheti, Vinay Aggarwal, Ananya Ganesh, Niyati Chhaya, and Nanda Kambhatla. 2022. Text Simplification for Legal Domain: Insights and Challenges. In *Proceedings of the Natural Legal Language Processing Workshop*, pages 296–304.

Nicolas Garneau, Eve Gaumond, Luc Lamontagne, and Pierre-Luc Déziel. 2021. CriminelBART: A French Canadian Legal Language Model Specialized in Criminal Law. In *Proceedings of the International Conference on Artificial Intelligence and Law*, ICAIL '21, page 256–257, New York, NY, USA. Association for Computing Machinery.

Nicolas Garneau, Eve Gaumond, Luc Lamontagne, and Pierre-Luc Déziel. 2022. Evaluating legal accuracy of neural generators on the generation of criminal court dockets description. In *Proceedings of the International Conference on Natural Language Generation*, pages 73–99, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Albert Gatt and Emiel Krahmer. 2018. Survey of the State of the Art in Natural Language Generation: Core Tasks, Applications and Evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Natalia Grabar and Rémi Cardon. 2018. Clear-Simple Corpus for Medical French. In *ATA*.

Margaret Hagan. 2023. Good AI Legal Help, Bad AI Legal Help: Establishing Quality Standards for Responses to People's Legal Problem Stories. In *JURIX*, volume 2023.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication methods and measures*, 1(1):77–89.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. SpaCy: Industrial-strength Natural Language Processing in Python.

Debra Howcroft and Birgitta Bergvall-Kåreborn. 2019. A Typology of Crowdwork Platforms. *Work, employment and society*, 33(1):21–38.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, et al. 2024. Improving Large Language Models for Clinical Named Entity Recognition via Prompt Engineering. *Journal of the American Medical Informatics Association*, page ocad259.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*, volume 112. Springer.

Samyar Janatian, Hannes Westermann, Jinzhe Tan, Jaromir Savelka, and Karim Benyekhlef. 2023. From Text to Structure: Using Large Language Models to Support the Development of Legal Expert Systems. *Legal Knowledge and Information Systems*, pages 167–176.

Hans Kamp and Uwe Reyle. 2013. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*, volume 42. Springer Science & Business Media.

Liliane Kandel and Abraham Moles. 1958. Application de l'indice de flesch à la langue française. *Cahiers Etudes de Radio-Télévision*, 19(1958):253–274.

Sayash Kapoor, Peter Henderson, and Arvind Narayanan. 2024. Promises and Pitfalls of Artificial Intelligence for Legal Applications. *Forthcoming in the Journal of Cross-disciplinary Research in Computational Law (CRCL)*.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael James Bommarito. 2023. Natural Language Processing in the Legal Domain. *Available at SSRN 4336224*.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking Large Language Models on Sentence Simplification. *arXiv:2310.15773*.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. Mitigating societal harms in large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 26–33, Singapore. Association for Computational Linguistics.

Philippe Laban, Andrew Hsi, John Canny, and Marti A Hearst. 2020. The Summary Loop: Learning to Write Abstractive Summaries Without Examples. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 5135–5150.

Ashley Lewis and Michael White. 2023. Mitigating harms of LLMs via knowledge distillation for a virtual museum tour guide. In *Proceedings of the Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants!*, pages 31–45, Prague, Czech Republic. Association for Computational Linguistics.

Rensis Likert. 1932. *A Technique for the Measurement of Attitudes*. Archives of psychology ; no. 140. [s.n.], New York.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2022. LENS: A Learnable Evaluation Metric for Text Simplification. *arXiv:2212.09739*.

Margot Madina, Itziar Gonzalez-Dios, and Melanie Siegel. 2024. A Preliminary Study of ChatGPT for Spanish E2R Text Adaptation. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 1422–1434.

Moffatt *v. Air Canada*. 2024. Civil Resolution Tribunal of British Columbia 149.

Ines Montani and Matthew Honnibal. 2018. Prodigy: A Modern and Scriptable Annotation Tool for Creating Training Data for Machine Learning Models.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the Stability of Fine-Tuning BERT: Misconceptions, Explanations, and Strong Baselines. In *International Conference on Learning Representations*.

Debora Nozza, Giuseppe Attanasio, et al. 2023. Is It Really That Simple? Prompting Language Models for Automatic Text Simplification in Italian. In *CEUR Workshop Proceedings*. (seleziona).

Sinan Ozdemir. 2023. *Quick Start Guide to Large Language Models: Strategies and Best Practices for Using ChatGPT and Other LLMs*. Addison-Wesley Professional.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the annual meeting of the Association for Computational Linguistics*, pages 311–318.

Vincent Primpied, David Beauchemin, and Richard Khoury. 2022. Quantifying French Document Complexity. *Proceedings of the Canadian Conference on Artificial Intelligence*. Https://caiac.pubpub.org/pub/iaeeogod.

Quebec. 2022a. Code de la sécurité routière.

Quebec. 2022b. Loi modifiant la loi sur l'assurance automobile, le code de la sécurité routière et d'autres dispositions.

Clément Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, and Patrick Gallinari. 2021. Data-QuestEval: A Referenceless Metric for Data-To-Text Semantic Evaluation. In *Conference on Empirical Methods in Natural Language Processing*, pages 8029–8036. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 3982–3992.

Michael J Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting Non-english Text Simplification: A Unified Multilingual Benchmark. *arXiv:2305.15678*.

Horacio Saggion. 2017. Automatic Text Simplification. *Synthesis Lectures on Human Language Technologies*, 10(1):1–137.

Anastasia Shimorina and Anya Belz. 2021. The Human Evaluation Datasheet 1.0: A Template for Recording Details of Human Evaluation Experiments in NLP. *arXiv:2103.09710*.

Sanja Štajner. 2021. Automatic Text Simplification for Social Good: Progress and Challenges. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652.

Regina Stodden. 2021. When the Scale is Unclear-Analysis of the Interpretation of Rating Scales in Human Evaluation of Text Simplification. In *CTTS@ SEPLN*.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is Not Suitable for the Evaluation of Text Simplification. In *Conference on Empirical Methods in Natural Language Processing*, pages 738–744. Association for Computational Linguistics.

Jinzhe Tan, Hannes Westermann, and Karim Benyekhlef. 2023. ChatGPT as an Artificial Lawyer. *Artificial Intelligence for Access to Justice*.

Wilson L Taylor. 1953. “Cloze Procedure”: A New Tool for Measuring Readability. *Journalism quarterly*, 30(4):415–433.

Petter Törnberg. 2024. Best Practices for Text Annotation with Large Language Models. *arXiv:2402.05129*.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Roeland Van Hout and Anne Vermeer. 2007. Comparing Measures of Lexical Richness. *Modelling and assessing vocabulary knowledge*, 93:115.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and Social Risks of Harm From Language Models. *arXiv:2112.04359*.

Xuanxin Wu and Yuki Arase. 2024. An In-depth Evaluation of GPT-4 in Sentence Simplification with Error-based Human Assessment. *arXiv:2403.04963*.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

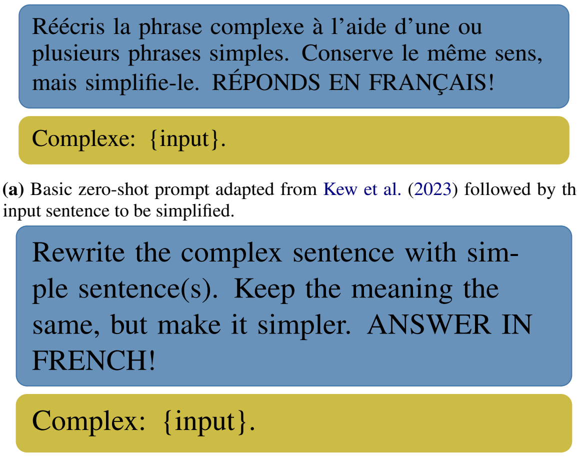
Jerrold H Zar. 2005. Spearman Rank Correlation. *Encyclopedia of Biostatistics*, 7.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation With BERT. In *International Conference on Learning Representations*.

A Automatic Text Simplification Prompt and Generation Parameters

The Figure 3a presents the French prompt used for generating the TS, along with the automatic translation in English using DeepL machine translator[8] (Figure 3b for the non-French reader); `{input}` is the placeholder for the complex sentence. It is based on Kew et al. (2023) basic zero-shot prompt, but we made the following two modifications:

1. Manual translation from English to French;
2. Add specification in uppercase to respond in French since uppercase capitalization has increased importance to an instruction (Ozdemir, 2023; Törnberg, 2024; Hu et al., 2024).



Réécris la phrase complexe à l’aide d’une ou plusieurs phrases simples. Conserve le même sens, mais simplifie-le. RÉPONDS EN FRANÇAIS!

Complexe: {input}.

(a) Basic zero-shot prompt adapted from Kew et al. (2023) followed by the input sentence to be simplified.

Rewrite the complex sentence with simple sentence(s). Keep the meaning the same, but make it simpler. ANSWER IN FRENCH!

Complex: {input}.

(b) Translation of the prompt presented in Figure 3a.

Figure 3: Prompts used for LLM text simplification. Bleu boxes contain the task instructions. Yellow boxes contain the prefix for the model to continue.

Table 7 present the OpenAI generation parameters used for generating the simplification. We have used the same parameters as per Kew et al. (2023). The cost of generating all 312 simplifications was less than 5 USD.

Parameter	Value
Model name	gpt-4-turbo-2024-04-09
Max new tokens	100
Temperature	1.0
Top K	0.9
Frequency penalty	0.0
Presence penalty	0.0

Table 7: OpenAI generation parameters used for generating the simplification

[8] https://www.deepl.com/translator

B Generation Examples

The Figure 4 presents examples of original text in French (cyan) along with the simplification (pink) made from the GPT4 model using the zero-shot simplification prompt and generation parameters as presented in Appendix A, and their respective automatic translation in English using DeepL machine translator (purple, brown respectively).

> Franchise, Il s'agit d'un montant restant à votre charge en cas de sinistre. Ce montant est stipulé aux Conditions particulières.

> Franchise est le montant que vous payez en cas de sinistre. Ce montant est indiqué dans les Conditions particulières.

> Deductible: This is the amount you must pay in the event of a claim. This amount is stipulated in the Declarations.

> Deductible is the amount you pay in the event of a claim. This amount is indicated in the Declarations.

(a) Example of a generation based on the deductible definition from the Quebec automotive insurance form.

> Aux conditions ci-après, l'alinéa de l'exclusion no 15 est supprimé en ce qui concerne la garantie A de la première partie - Assurance de vos biens, de votre contrat.

> L'alinea de l'exclusion no 15 est supprimé de la garantie A de la première partie - Assurance de vos biens, de votre contrat, sous certaines conditions.

> Under the following conditions, the paragraph of exclusion no. 15 is deleted with respect to Coverage A of Part I - Property Insurance of your contract.

> Under certain conditions, the paragraph of exclusion no. 15 is deleted from coverage A of the first part of your contract - Insurance of your property.

(b) Example of a generation based on the deductible definition from the Quebec principal residency insurance form.

Figure 4: Examples of a generation using GPT4. Cyan boxes contain the original French text, and purple boxes contain the automatic translation of the original text in English. Pink boxes contain the simplification generation in French, and brown boxes contain the automatic translation of the simplified generation in English.

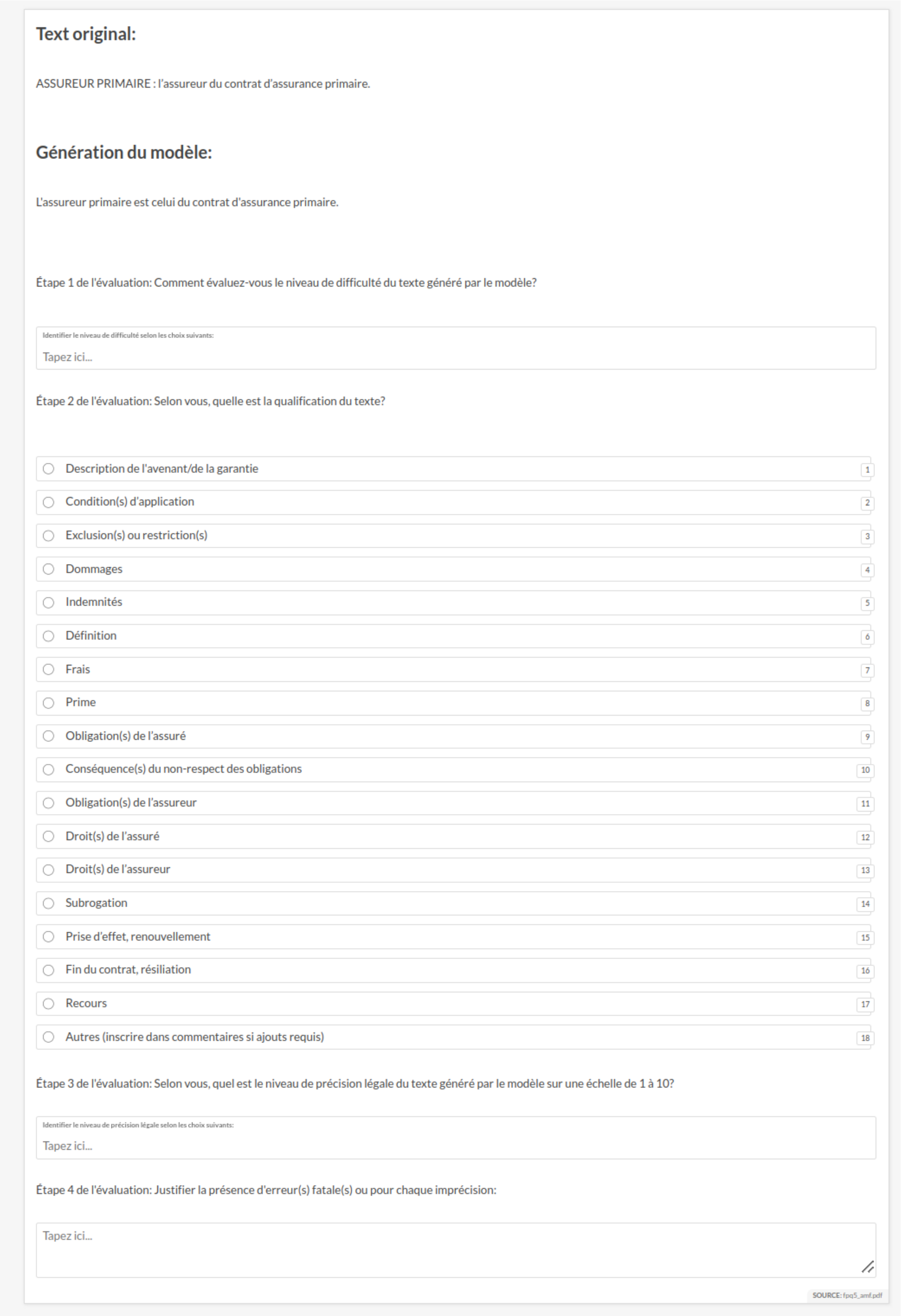
C Annotation Interface

The Figure 5 presents the evaluation interface used by our annotators (in French). It is a custom adaptation of the Prodigy annotation tool (Montani and Honnibal, 2018).

D Characterization Class

In this section, we detail the characterization class used by our annotator. For each, we present the characterization in French, an automatic English translation, and a brief description in English. All description were taken from Caron (2024).

1. **Description of endorsement (*Description de l'avenant*)**: These are appendices that modify the basic insurance contract, such as the "replacement cost" coverage endorsement. The text of the endorsement takes precedence over the general text of the insurance policy.
2. **Conditions of application (*Conditions d'applications*)**: Refers to the general conditions of application of either an insurance contract or endorsements. For example, "subject to risk acceptance".
3. **Exclusions or restrictions (*Exclusions ou restrictions*)**: Refers to the general exclusions or restrictions that can apply to the insurance contract or the endorsements. For example, "exclusions of replacement value" or "exclusions of nuclear damage".
4. **Damage (value of, calculation of and description of) (*Dommages (valeur des, calcul des et description des)*)**: Refers to the mechanism and principles to assess the value of the damage after an incident.
5. **Indemnities (indemnities payable, indemnity per replacement, calculation of value of, amount of insurance and indemnity process) (*Indemnités (indemnités payables, indemnité par remplacement, calcul de la valeur des, montant d'assurance et processus d'indemnisation)*)**: Refers to the mechanism and principles to assess the indemnities amount payable to an insuree, the principles of a replacement of the damaged property, the methodology to evaluate the value of the damage properties and indemnisation process along with the resolution in case of disagreement.
6. **Definition (*Définition*)**: Refers to definitions of specific terms in the contract, endorsements or other legal elements. For example, "definition



Text original:

ASSUREUR PRIMAIRE : l'assureur du contrat d'assurance primaire.

Génération du modèle:

L'assureur primaire est celui du contrat d'assurance primaire.

Étape 1 de l'évaluation: Comment évaluez-vous le niveau de difficulté du texte généré par le modèle?

Identifier le niveau de difficulté selon les choix suivants:

Tapez ici...

Étape 2 de l'évaluation: Selon vous, quelle est la qualification du texte?

- ○ Description de l'avenant/de la garantie 1
- ○ Condition(s) d'application 2
- ○ Exclusion(s) ou restriction(s) 3
- ○ Dommages 4
- ○ Indemnités 5
- ○ Définition 6
- ○ Frais 7
- ○ Prime 8
- ○ Obligation(s) de l'assuré 9
- ○ Conséquence(s) du non-respect des obligations 10
- ○ Obligation(s) de l'assureur 11
- ○ Droit(s) de l'assuré 12
- ○ Droit(s) de l'assureur 13
- ○ Subrogation 14
- ○ Prise d'effet, renouvellement 15
- ○ Fin du contrat, résiliation 16
- ○ Recours 17
- ○ Autres (inscrire dans commentaires si ajouts requis) 18

Étape 3 de l'évaluation: Selon vous, quel est le niveau de précision légale du texte généré par le modèle sur une échelle de 1 à 10?

Identifier le niveau de précision légale selon les choix suivants:

Tapez ici...

Étape 4 de l'évaluation: Justifier la présence d'erreur(s) fatale(s) ou pour chaque imprécision:

Tapez ici...

SOURCE: fpq5_amf.pdf

Figure 5: The Prodigy annotation interface (in French) used by the annotators to evaluate the instance generated by an ATS system.

of deductible”.

7. **Expenses (reimbursement and assumption of costs) (*Frais (remboursement et prise en charge des)*)**: Refers to the principles of expense reimbursement in case of an insured incident, such as towing the insured car or expenses to minimize damage.
8. **Premium (payment and reimbursement of) (*Prime (paiement de et remboursement de)*)**: Refers to premium details such as the amount, how and when to pay it, and the reimbursement terms.
9. **Obligations of the insured (obligation and formal commitment) (*Obligations de l’assuré (obligation et engagement formel)*)**: Refers to the insuree’s obligations to be executed during the duration of the contract. For example, “Risk aggravation declaration”.
10. **Consequences of non-compliance (*Conséquences du non-respect des obligations*)**: Refers to the consequences of non-compliance to the insuree or insurer engagements, such as indemnity reduction or legal actions of the insurer against its insuree (e.g. false declaration).
11. **Insurer’s obligations (*Obligations de l’assureur*)**: Refers to the insurer’s obligations to be executed during the contract duration. For example, “insurer’s obligation to inform and advise the insured”.
12. **Insured’s rights (including waiver of rights) (*Droits de l’assuré (incluant la renonciation aux droits)*)**: Refers to the rights of the insured regarding the insurance contract, such as the right of renewal and representation.
13. **Insurer’s rights (including waiver of rights) (*Droits de l’assureur (incluant la renonciation aux droits)*)**: Refers to the insurer’s right regarding the insurance contract, such as the right to refuse coverage.
14. **Subrogation (and exceptions to subrogation) (*Subrogation (et exceptions à la subrogation)*)**: Refers to a specific right of the insuree and insured called subrogation right that defines the right of the insuree to transfer all its rights over an incident to the insurer. The insurer will represent the right of the insuree and protect the insuree and insurer interest.
15. **Effective date and renewal (*Prise d’effet et renouvellement*)**: Refers to the effective date and renewal of the insurance contract.
16. **End of contract and termination (*Fin du contrat et résiliation*)**: Refers to the effective end date and the termination conditions of the insurance contract.
17. **Legal recourse (dispute resolution, action, representation mandate, arbitration, etc.) (*Recours (règlement de différend, action, mandat de représentation, arbitrage, etc.)*)**: Refers to the mechanisms for resolving legal disputes and institutions to which policyholders can refer. For example, to the regulatory body (i.e. AMF in Quebec).
18. **Others (*Autres*)**: Class use when the sentence does not apply to any of the 17 previous classes. When annotators use these cases, we ask them to elaborate on why none of the earlier classes were appropriate.

E Training Loss

In this section, we present the training and evaluation loss for trained metrics in Figure 6. We can see in Figure 6a that the loss reaches a plateau after 60 epochs, resulting in a wide gap between the training and evaluation loss. It indicates potential overfitting for the JUDGEBERT model. However, as shown in Figure 6b, JUDGEBERT-DA training and evaluation gap is smaller and tends to slowly decrease over time, meaning a lower risk of overfitting.

F Human Evaluation Datasheet

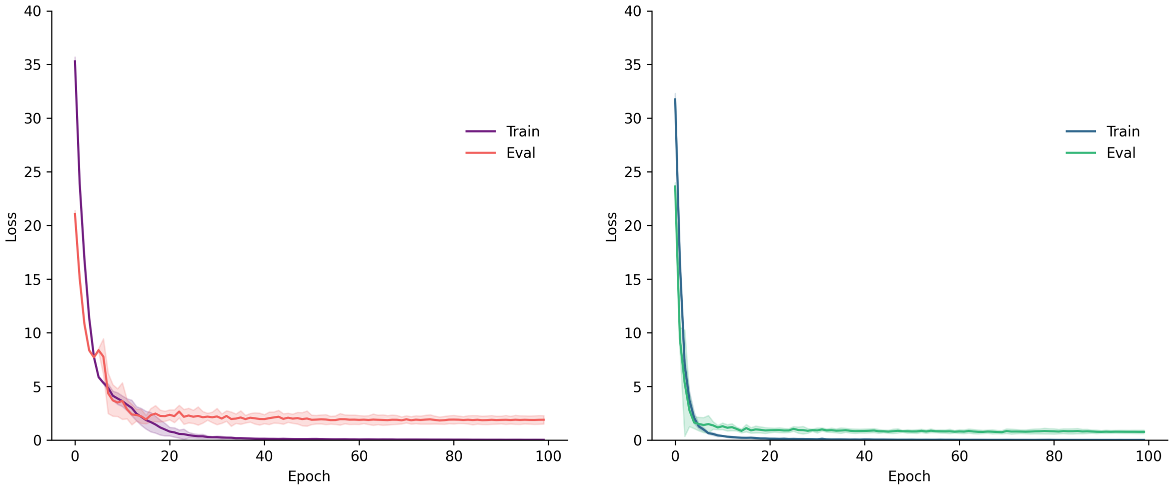
F.1 Paper and Supplementary Resources (Questions 1.1–1.3)

Question 1.1: Link to paper reporting the evaluation experiment. If the paper reports more than one experiment, state which experiment you’re completing this sheet for. Or, if applicable, enter ‘for preregistration.’

For preregistration.

Question 1.2: Link to website providing resources used in the evaluation experiment (e.g. system outputs, evaluation tools, etc.). If there isn’t one, enter ‘N/A’.

N/A.



(a) Training and evaluation loss for `JUDGEBERT` over the 100 epochs.

(b) Training and evaluation loss for `JUDGEBERT-DA` over the 100 epochs.

Figure 6: Training and evaluation loss

Question 1.3: Name, affiliation and email address of person completing this sheet, and of contact author if different.

David Beauchemin
david.beauchemin@ift.ulaval.ca

F.2 System (Questions 2.1–2.5)

Question 2.1: What type of input do the evaluated system(s) take? Select all that apply. If none match, select 'Other' and describe.

Check-box options (select all that apply):

✓ ***raw/structured data***: numerical, symbolic, and other data, possibly structured into trees, graphs, graphical models, etc. May be the input e.g. to Referring Expression Generation (REG), end-to-end text generation, etc. NB: excludes linguistic structures.

□ ***deep linguistic representation (DLR)***: any of a variety of deep, underspecified, semantic representations, such as abstract meaning representations (AMRs; Banarescu et al., 2013) or discourse representation structures (DRSs; Kamp and Reyle, 2013).

□ ***shallow linguistic representation (SLR)***: any of a variety of shallow, syntactic representations, e.g. Universal Dependency (UD) structures; typically the input to surface realisation.

□ ***text: subsentential unit of text***: a unit of text shorter than a sentence, e.g. Referring Expressions (REs), verb phrase, text fragment of any length; includes titles/headlines.

□ ***text: sentence***: a single sentence (or set of sentences).

✓ ***text: multiple sentences***: a sequence of multiple sentences, without any document structure (or a set of such sequences).

□ ***text: document***: a text with document structure, such as a title, paragraph breaks or sections, e.g. a set of news reports for summarisation.

□ ***text: dialogue***: a dialogue of any length, excluding a single turn which would come under one of the other text types.

□ ***text: other***: input is text but doesn't match any of the above *text:** categories.

□ ***speech***: a recording of speech.

□ ***visual***: an image or video.

□ ***multi-modal***: catch-all value for any combination of data and/or linguistic representation and/or visual data etc.

□ ***control feature***: a feature or parameter specifically present to control a property of the output text, e.g. positive stance, formality, author style.

□ ***no input (human generation)***: human generation[9], therefore no system inputs.

□ ***other (please specify)***: if input is none of the above, choose this option and describe it.

[9] We use the term 'human generation' where the items being evaluated have been created manually, rather than generated by an automatic system.

Question 2.2: What type of output do the evaluated system(s) generate? Select all that apply. If none match, select 'Other' and describe.

Check-box options (select all that apply):

□ ***raw/structured data***: numerical, symbolic, and other data, possibly structured into trees, graphs, graphical models, etc. May be the input e.g. to Referring Expression Generation (REG), end-to-end text generation, etc. NB: excludes linguistic structures.

□ ***deep linguistic representation (DLR)***: any of a variety of deep, underspecified, semantic representations, such as abstract meaning representations (AMRs; Banarescu et al., 2013) or discourse representation structures (DRSs; Kamp and Reyle, 2013).

□ ***shallow linguistic representation (SLR)***: any of a variety of shallow, syntactic representations, e.g. Universal Dependency (UD) structures; typically the input to surface realisation.

□ ***text: subsentential unit of text***: a unit of text shorter than a sentence, e.g. Referring Expressions (REs), verb phrase, text fragment of any length; includes titles/headlines.

□ ***text: sentence***: a single sentence (or set of sentences).

✓ ***text: multiple sentences***: a sequence of multiple sentences, without any document structure (or a set of such sequences).

□ ***text: document***: a text with document structure, such as a title, paragraph breaks or sections, e.g. a set of news reports for summarisation.

□ ***text: dialogue***: a dialogue of any length, excluding a single turn which would come under one of the other text types.

□ ***text: other***: select if output is text but doesn't match any of the above *text:** categories.

□ ***speech***: a recording of speech.

□ ***visual***: an image or video.

□ ***multi-modal***: catch-all value for any combination of data and/or linguistic representation and/or visual data etc.

□ ***human-generated 'outputs'***: manually created stand-ins exemplifying outputs.

□ ***other (please specify)***: if output is none of the above, choose this option and describe it.

Question 2.3: How would you describe the task that the evaluated system(s) perform in mapping the inputs in Q2.1 to the outputs in Q2.2? Occasionally, more than one of the options below may apply. If none match, select 'Other' and describe.

Check-box options (select all that apply):

□ ***content selection/determination***: selecting the specific content that will be expressed in the generated text from a representation of possible content. This could be attribute selection for REG (without the surface realisation step). Note that the output here is not text.

□ ***content ordering/structuring***: assigning an order and/or structure to content to be included in generated text. Note that the output here is not text.

□ ***aggregation***: converting inputs (typically *deep linguistic representations* or *shallow linguistic representations*) in some way in order to reduce redundancy (e.g. representations for 'they like swimming', 'they like running' → representation for 'they like swimming and running').

□ ***referring expression generation***: generating *text* to refer to a given referent, typically represented in the input as a set of attributes or a linguistic representation.

□ ***lexicalisation***: associating (parts of) an input representation with specific lexical items to be used in their realisation.

✓ ***deep generation***: one-step text generation from *raw/structured data* or *deep linguistic representations*. One-step means that no intermediate representations are passed from one independently run module to another.

□ ***surface realisation (SLR to text)***: one-step text generation from *shallow linguistic representations*. One-step means that no intermediate representations are passed from one independently run module to another.

□ ***feature-controlled text generation***: generation of text that varies along specific dimensions where the variation is controlled via *control features* specified as part of the input. Input is a non-textual representation (for feature-controlled text-to-text generation select the matching text-to-text task).

□ ***data-to-text generation***: generation from *raw/structured data* which may or may not include some amount of content selection as part of the generation process. Output is likely to be *text:** or *multi-modal*.

□ ***dialogue turn generation***: generating a dialogue turn (can be a greeting or closing) from a representation of dialogue state and/or last turn(s), etc.

□ ***question generation***: generation of questions from given input text and/or knowledge base such that the question can be answered from the input.

□ ***question answering***: input is a question plus optionally a set of reference texts and/or knowledge base, and the output is the answer to the question.

✓ ***paraphrasing/lossless simplification***: text-to-text generation where the aim is to preserve the meaning of the input while changing its wording. This can include the aim of changing the text on a given dimension, e.g. making it simpler, changing its stance or sentiment, etc., which may be controllable via input features. Note that this task type includes meaning-preserving text simplification (non-meaning preserving simplification comes under *compression/lossy simplification* below).

□ ***compression/lossy simplification***: text-to-text generation that has the aim to generate a shorter, or shorter and simpler, version of the input text. This will normally affect meaning to some extent, but as a side effect, rather than the primary aim, as is the case in *summarisation*.

□ ***machine translation***: translating text in a source language to text in a target language while maximally preserving the meaning.

□ ***summarisation (text-to-text)***: output is an extractive or abstractive summary of the important/relevant/salient content of the input document(s).

□ ***end-to-end text generation***: use this option if the single system task corresponds to more than one of tasks above, implemented either as separate modules pipelined together, or as one-step generation, other than *deep generation* and *surface realisation*.

□ ***image/video description***: input includes *visual*, and the output describes it in some way.

□ ***post-editing/correction***: system edits and/or corrects the input text (typically itself the textual output from another system) to yield an improved version of the text.

□ ***other (please specify)***: if task is none of the above, choose this option and describe it.

Question 2.4: Input Language(s), or 'N/A'.

French.

Question 2.5: Output Language(s), or 'N/A'.

French.

F.3 Output Sample, Evaluators, Experimental Design

F.3.1 Sample of system outputs (or human-authored stand-ins) evaluated (Questions 3.1.1–3.1.3)

Question 3.1.1: How many system outputs (or other evaluation items) are evaluated per system in the evaluation experiment? Answer should be an integer.

297.

Question 3.1.2: How are system outputs (or other evaluation items) selected for inclusion in the evaluation experiment? If none match, select 'Other' and describe.

Multiple-choice options (select one):

○ ***by an automatic random process from a larger set***: outputs were selected for inclusion in the experiment by a script using a pseudo-random number generator; don't use this option if the script selects every nth output (which is not random).

○ ***by an automatic random process but using stratified sampling over given properties***: use this option if selection was by a random script as above, but with added constraints ensuring that the sample is representative of the set of outputs it was selected from, in terms of given properties, such as sentence length, positive/negative stance, etc.

○ ***by manual, arbitrary selection***: output sample was selected by hand, or automatically from a manually compiled list, without a specific selection criterion.

✓ ***by manual selection aimed at achieving balance or variety relative to given properties***: selection by hand as above, but with specific selection criteria, e.g. same number of outputs from each time period.

○ ***Other (please specify)***: if selection method is none of the above, choose this option and describe it.

Question 3.1.3: What is the statistical power of the sample size?

Following the methodology of Card et al. (2020), we obtained a statistical power of 0.33 on the output sample w.r.t the automatic evaluation metrics, the two best-performing models (`JUDGEBERT-DA` and BERTScore). We used their online script to estimate the statistical power.

F.3.2 Evaluators (Questions 3.2.1–3.2.4)

Question 3.2.1: How many evaluators are there in this experiment? Answer should be an integer.

Five.

Question 3.2.2: What kind of evaluators are in this experiment? Select all that apply. If none match, select 'Other' and describe. In all cases, provide details in the text box under 'Other'.

Check-box options (select all that apply):

✓ ***experts***: participants are considered domain experts, e.g. meteorologists evaluating a weather forecast generator, or nurses evaluating an ICU report generator.

□ ***non-experts***: participants are not domain experts.

✓ ***paid (including non-monetary compensation such as course credits)***: participants were given some form of compensation for their participation, including vouchers, course credits, and reimbursement for travel unless based on receipts.

□ ***not paid***: participants were not given compensation of any kind.

□ ***previously known to authors***: (one of the) researchers running the experiment knew some or all of the participants before recruiting them for the experiment.

✓ ***not previously known to authors***: none of the researchers running the experiment knew any of the participants before recruiting them for the experiment.

□ ***evaluators include one or more of the authors***: one or more researchers running the experiment was among the participants.

✓ ***evaluators do not include any of the authors***: none of the researchers running the experiment were among the participants.

□ ***Other*** (fewer than 4 of the above apply): we believe you should be able to tick 4 options of the above. If that's not the case, use this box to explain.

Question 3.2.3: How are evaluators recruited?

Evaluators were recruited through a job offer on the University job board and interviewed prior to conducting the experiment.

Question 3.2.4: What training and/or practice are evaluators given before starting on the evaluation itself?

First, the evaluators have been introduced to the task of text simplification generation. They were then introduced to the dataset under study. They learned from an annotation guideline and practices on 15 examples before conducting the whole experiment. Evaluators did not need legal training since they all had domain background knowledge.

Question 3.2.5: What other characteristics do the evaluators have, known either because these were qualifying criteria, or from information gathered as part of the evaluation?

Evaluators have been selected based on their educational level, i.e. at least in their second year in law school, and interest in insurance law.

F.3.3 Experimental design (Questions 3.3.1–3.3.8)

Question 3.3.1: Has the experimental design been preregistered? If yes, on which registry?

No.

Question 3.3.2: How are responses collected? E.g. paper forms, online survey tool, etc.

The answers were collected using a customized version of Prodigy[10], hosted on Amazon Web Services.

Question 3.3.3: What quality assurance methods are used? Select all that apply. If none match, select 'Other' and describe. In all cases, provide details in the text box under 'Other'.

Check-box options (select all that apply):

✓ ***evaluators are required to be native speakers of the language they evaluate***: mechanisms are in place to ensure all participants are native speakers of the language they evaluate.

□ ***automatic quality checking methods are used during/post evaluation***: evaluations are checked for quality by automatic scripts during or after evaluations, e.g. evaluators are given known bad/good outputs to check they're given bad/good scores on MTurk.

✓ ***manual quality checking methods are used during/post evaluation***: evaluations are checked for quality by a manual process during or after evaluations, e.g. scores assigned by evaluators are monitored by researchers conducting the experiment.

□ ***evaluators are excluded if they fail quality checks (often or badly enough)***: there are conditions under which evaluations produced by participants are not included in the final results due to quality issues.

□ ***some evaluations are excluded because of failed quality checks***: there are conditions under which some (but not all) of the evaluations produced by some participants are not included in the final results due to quality issues.

□ ***none of the above***: tick this box if none of the above apply.

□ ***Other (please specify)***: use this box to describe any other quality assurance methods used during or after evaluations, and to provide additional details for any of the options selected above.

Question 3.3.4: What do evaluators see when carrying out evaluations? Link to screenshot(s) and/or describe the evaluation interface(s).

When evaluating, evaluators see the input data (e.g., a complex sentence) and the simplification generated by the model. To reduce any bias toward public LLM (e.g., GPT4), they do not know the model name. They then independently provide a score for each generation.

3.3.5: How free are evaluators regarding when and how quickly to carry out evaluations? Select all that apply. In all cases, provide details in the text box under 'Other'.

[10] https://prodi.gy/

Check-box options (select all that apply):

□ ***evaluators have to complete each individual assessment within a set time***: evaluators are timed while carrying out each assessment and cannot complete the assessment once time has run out.

□ ***evaluators have to complete the whole evaluation in one sitting***: partial progress cannot be saved and the evaluation returned to on a later occasion.

✓ ***neither of the above***: Choose this option if neither of the above are the case in the experiment.

□ ***Other (please specify)***: Use this space to describe any other way in which time taken or number of sessions used by evaluators is controlled in the experiment, and to provide additional details for any of the options selected above.

3.3.6: Are evaluators told they can ask questions about the evaluation and/or provide feedback? Select all that apply. In all cases, provide details in the text box under 'Other'.

Check-box options (select all that apply):

✓ ***evaluators are told they can ask any questions during/after receiving initial training/instructions, and before the start of the evaluation***: evaluators are told explicitly that they can ask questions about the evaluation experiment *before* starting on their assessments, either during or after training.

□ ***evaluators are told they can ask any questions during the evaluation***: evaluators are told explicitly that they can ask questions about the evaluation experiment *during* their assessments.

□ ***evaluators are asked for feedback and/or comments after the evaluation, e.g. via an exit questionnaire or a comment box***: evaluators are explicitly asked to provide feedback and/or comments about the experiment *after* their assessments, either verbally or in written form.

□ ***None of the above***: Choose this option if none of the above are the case in the experiment.

□ ***Other (please specify)***: use this space to describe any other ways you provide for evaluators to ask questions or provide feedback.

3.3.7: What are the experimental conditions in which evaluators carry out the evaluations? If none match, select 'Other' and describe.

Multiple-choice options (select one):

✓ ***evaluation carried out by evaluators at a place of their own choosing, e.g. online, using a paper form, etc.***: evaluators are given access to the tool or form specified in Question 3.3.2, and subsequently choose where to carry out their evaluations.

○ ***evaluation carried out in a lab, and conditions are the same for each evaluator***: evaluations are carried out in a lab, and conditions in which evaluations are carried out *are* controlled to be the same, i.e. the different evaluators all carry out the evaluations in identical conditions of quietness, same type of computer, same room, etc. Note we're not after very fine-grained differences here, such as time of day or temperature, but the line is difficult to draw, so some judgment is involved here.

○ ***evaluation carried out in a lab, and conditions vary for different evaluators***: choose this option if evaluations are carried out in a lab, but the preceding option does not apply, i.e. conditions in which evaluations are carried out are *not* controlled to be the same.

○ ***evaluation carried out in a real-life situation, and conditions are the same for each evaluator***: evaluations are carried out in a real-life situation, i.e. one that would occur whether or not the evaluation was carried out (e.g. evaluating a dialogue system deployed in a live chat function on a website), and conditions in which evaluations are carried out *are* controlled to be the same.

○ ***evaluation carried out in a real-life situation, and conditions vary for different evaluators***: choose this option if evaluations are carried out in a real-life situation, but the preceding option does not apply, i.e. conditions in which evaluations are carried out are *not* controlled to be the same.

○ ***evaluation carried out outside of the lab, in a situation designed to resemble a real-life situation, and conditions are the same for each*** *evaluator*: evaluations are carried out outside of the lab, in a situation intentionally similar to a real-life situation (but not actually a real-life situation), e.g. user-testing a navigation system where the destination is part of the evaluation design, rather than chosen by the user. Conditions in which evaluations are carried out *are* controlled to be the same.

○ ***evaluation carried out outside of the lab, in a situation designed to resemble a real-life situation, and conditions vary for different evaluators***: choose this option if evaluations are carried out outside of the lab, in a situation intentionally similar to a real-life situation, but the preceding option does not apply, i.e. conditions in which evaluations are carried out are *not* controlled to be the same.

○ ***Other (please specify)***: Use this space to provide additional, or alternative, information about the conditions in which evaluators carry out assessments, not covered by the options above.

3.3.8: Unless the evaluation is carried out at a place of the evaluators' own choosing, briefly describe the (range of different) conditions in which evaluators carry out the evaluations.

N/A.

F.4 Quality Criterion *n* – Definition and Operationalisation

F.4.1 Quality criterion properties (Questions 4.1.1–4.1.3)

Question 4.1.1: What type of quality is assessed by the quality criterion?

Multiple-choice options (select one):

✓ ***Correctness***: select this option if it is possible to state, generally for all outputs, the conditions under which outputs are maximally correct (hence of maximal quality). E.g. for Grammaticality, outputs are (maximally) correct if they contain no grammatical errors; for Semantic Completeness, outputs are correct if they express all the content in the input.

○ ***Goodness***: select this option if, in contrast to correctness criteria, there is no single, general mechanism for deciding when outputs are maximally good, only for deciding for two outputs which is better and which is worse. E.g. for Fluency, even if outputs contain no disfluencies, there may be other ways in which any given output could be more fluent.

○ ***Features***: choose this option if, in terms of property X captured by the criterion, outputs are not generally better if they are more X, but instead, depending on evaluation context, more X may be better or less X may be better. E.g. outputs can be more specific or less specific, but it's not the case that outputs are, in the general case, better when they are more specific.

Question 4.1.2: Which aspect of system outputs is assessed by the quality criterion?

Multiple-choice options (select one):

○ ***Form of output***: choose this option if the criterion assesses the form of outputs alone, e.g. Grammaticality is only about the form, a sentence can be grammatical yet be wrong or nonsensical in terms of content.

✓ ***Content of output***: choose this option if the criterion assesses the content/meaning of the output alone, e.g. Meaning Preservation only assesses output content; two sentences can be considered to have the same meaning, but differ in form.

○ ***Both form and content of output***: choose this option if the criterion assesses outputs as a whole, not just form or just content. E.g. Coherence is a property of outputs as a whole, either form or meaning can detract from it.

Question 4.1.3: Is each output assessed for quality in its own right, or with reference to a system-internal or external frame of reference?

Multiple-choice options (select one):

○ ***Quality of output in its own right***: choose this option if output quality is assessed without referring to anything other than the output itself, i.e. no system-internal or external frame of reference. E.g. Poeticness is assessed by considering (just) the output and how poetic it is.

✓ ***Quality of output relative to the input***: choose this option if output quality is assessed relative to the input. E.g. Answerability is the degree to which the output question can be answered from information in the input.

○ ***Quality of output relative to a system-external frame of reference***: choose this option if output quality is assessed with reference to system-external information, such as a knowledge base, a person's individual writing style, or the performance of an embedding system. E.g. Factual Accuracy assesses outputs relative to a source of real-world knowledge.

F.4.2 Evaluation mode properties (Questions 4.2.1–4.2.3)

Questions 4.2.1–4.2.3 record properties that are orthogonal to quality criteria, i.e. any given quality criterion can in principle be combined with any of the modes (although some combinations are more common than others).

Question 4.2.1: Does an individual assessment involve an objective or a subjective judgment?

Multiple-choice options (select one):

✓ ***Objective***: Examples of objective assessment include any automatically counted or otherwise quantified measurements such as mouse-clicks, occurrences in text, etc. Repeated assessments of the same output with an objective-mode evaluation method always yield the same score/result.

○ ***Subjective***: Subjective assessments involve ratings, opinions and preferences by evaluators. Some criteria lend themselves more readily to subjective assessments, e.g. Friendliness of a conversational agent, but an objective measure e.g. based on lexical markers is also conceivable.

Question 4.2.2: Are outputs assessed in absolute or relative terms?

Multiple-choice options (select one):

✓ ***Absolute***: choose this option if evaluators are shown outputs from a single system during each individual assessment.

○ ***Relative***: choose this option if evaluators are shown outputs from multiple systems at the same time during assessments, typically ranking or preference-judging them.

Question 4.2.3: Is the evaluation intrinsic or extrinsic?

Multiple-choice options (select one):

○ ***Intrinsic***: Choose this option if quality of outputs is assessed *without* considering their *effect* on something external to the system, e.g. the performance of an embedding system or of a user at a task.

✓ ***Extrinsic***: Choose this option if quality of outputs is assessed in terms of their *effect* on something external to the system such as the performance of an embedding system or of a user at a task.

F.4.3 Response elicitation (Questions 4.3.1–4.3.11)

Question 4.3.1: What do you call the quality criterion in explanations/interfaces to evaluators? Enter 'N/A' if criterion not named.

Legal meaning.

Question 4.3.2: What definition do you give for the quality criterion in explanations/interfaces to evaluators? Enter 'N/A' if no definition given.

We define **legal meaning** as "[the] measures [of] how well the output text conveys the legal details and exceptions and does not misrepresent the law".

Question 4.3.3: Size of scale or other rating instrument (i.e. how many different possible values there are). Answer should be an integer or 'continuous' (if it's not possible to state how many possible responses there are). Enter 'N/A' if there is no rating instrument.

10.

Question 4.3.4: List or range of possible values of the scale or other rating instrument. Enter 'N/A', if there is no rating instrument.

1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

Question 4.3.5: How is the scale or other rating instrument presented to evaluators? If none match, select 'Other' and describe.

Multiple-choice options (select one):

○ ***Multiple-choice options***: choose this option if evaluators select exactly one of multiple options.

○ ***Check-boxes***: choose this option if evaluators select any number of options from multiple given options.

○ ***Slider***: choose this option if evaluators move a pointer on a slider scale to the position corresponding to their assessment.

○ ***N/A (there is no rating instrument)***: choose this option if there is no rating instrument.

✓ ***Other (please specify)***: choose this option if there is a rating instrument, but none of the above adequately describe the way you present it to evaluators. Use the text box to describe the rating instrument and link to a screenshot.

Due to Prodigy's limitations regarding their slider component (only one per page), we used a free-form text box. Since we have few highly skilled evaluators, collecting data was not a problem.

Question 4.3.6: If there is no rating instrument, describe briefly what task the evaluators perform (e.g. ranking multiple outputs, finding information, playing a game, etc.), and what information is recorded. Enter 'N/A' if there is a rating instrument.

N/A.

Question 4.3.7: What is the verbatim question, prompt or instruction given to evaluators (visible to them during each individual assessment)?

Here is the verbatim question and instruction in French to evaluators (in the following list), we also present an automatic translation of these instruction in the second list.

Étape 1 de l'évaluation: *Comment évaluez-vous le niveau de difficulté du texte généré par le modèle?*
Étape 2 de l'évaluation: *Selon vous, quelle est la qualification du texte?*
Étape 3 de l'évaluation: *Selon vous, quel est le niveau de précision légale du texte généré par le modèle sur une échelle de 1 à 10?*
Commentaires (si applicable):

Here is the automatic English translation of the verbatim question and instructions for evaluators.

Evaluation step 1: How would you rate the level of difficulty of the text generated by the template?
Evaluation step 2: How do you rate the text?
Evaluation step 3: What do you think is the legal accuracy level of the text generated by the model on a scale of 1 to 10?
Comments (if applicable):

Question 4.3.8: Form of response elicitation. If none match, select 'Other' and describe.

Multiple-choice options (select one):[11]

[11] Explanations adapted from Howcroft and Bergvall-Kåreborn (2019).

○ ***(dis)agreement with quality statement***: Participants specify the degree to which they agree with a given quality statement by indicating their agreement on a rating instrument. The rating instrument is labelled with degrees of agreement and can additionally have numerical labels. E.g. *This text is fluent — 1=strongly disagree...5=strongly agree*.

○ ***direct quality estimation***: Participants are asked to provide a rating using a rating instrument, which typically (but not always) mentions the quality criterion explicitly. E.g. *How fluent is this text? — 1=not at all fluent...5=very fluent*.

○ ***relative quality estimation (including ranking)***: Participants evaluate two or more items in terms of which is better. E.g. *Rank these texts in terms of fluency*; *Which of these texts is more fluent?*; *Which of these items do you prefer?*.

✓ ***counting occurrences in text***: Evaluators are asked to count how many times some type of phenomenon occurs, e.g. the number of facts contained in the output that are inconsistent with the input.

○ ***qualitative feedback (e.g. via comments entered in a text box)***: Typically, these are responses to open-ended questions in a survey or interview.

○ ***evaluation through post-editing/annotation***: Choose this option if the evaluators' task consists of editing or inserting annotations in text. E.g. evaluators may perform error correction and edits are then automatically measured to yield a numerical score.

○ ***output classification or labelling***: Choose this option if evaluators assign outputs to categories. E.g. *What is the overall sentiment of this piece of text? — Positive/neutral/negative.*

○ ***user-text interaction measurements***: choose this option if participants in the evaluation experiment interact with a text in some way, and measurements are taken of their interaction. E.g. reading speed, eye movement tracking, comprehension questions, etc. Excludes situations where participants are given a task to solve and their performance is measured which comes under the next option.

○ ***task performance measurements***: choose this option if participants in the evaluation experiment are given a task to perform, and measurements are taken of their performance at the task. E.g. task is finding information, and task performance measurement is task completion speed and success rate.

○ ***user-system interaction measurements***: choose this option if participants in the evaluation experiment interact with a system in some way, while measurements are taken of their interaction. E.g. duration of interaction, hyperlinks followed, number of likes, or completed sales.

○ ***Other (please specify)***: Use the text box to describe the form of response elicitation used in assessing the quality criterion if it doesn't fall in any of the above categories.

Question 4.3.9: How are raw responses from participants aggregated or otherwise processed to obtain reported scores for this quality criterion? State if no scores reported.

Macro averages are computed from numerical scores to provide a summary.

Question 4.3.10: Method(s) used for determining effect size and significance of findings for this quality criterion.

What to enter in the text box: A list of methods used for calculating the effect size and significance of any results, both as reported in the paper given in Question 1.1, for this quality criterion. If none calculated, state 'None'.

None.

Question 4.3.11: Has the inter-annotator and intra-annotator agreement between evaluators for this quality criterion been measured? If yes, what method was used, and what are the agreement scores?

Krippendorff's alpha (Hayes and Krippendorff, 2007) is used to measure inter-annotator agreement. Krippendorff's alpha are detailled in Table 2.

F.5 Ethics

Question 5.1: Has the evaluation experiment this sheet is being completed for, or the larger study it is part of, been approved by a research ethics committee? If yes, which research ethics committee?

No.

Question 5.2: Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain personal data (as defined in GDPR Art. 4, §1: https://gdpr.eu/article-4-definitions/)? If yes, describe data and state how addressed.

No.

Question 5.3: Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain special category information (as defined in GDPR Art. 9, §1: https://gdpr.eu/article-9-processing-special-categories-of-personal-data-prohibited/)? If yes, describe data and state how addressed.

No.

Question 5.4: Have any impact assessments been carried out for the evaluation experiment, and/or any data collected/evaluated in connection with it? If yes, summarise approach(es) and outcomes.

No.